Differences in Outcomes Reported by Patients With Inflammatory Bowel Diseases vs Their Health Care Professionals

Pittet, Valérie E H ; Maillard, Michel H ; Simonson, Thomas ; Fournier, Nicolas ; Rogler, Gerhard ; Michetti, Pierre ; Swiss IBD Cohort Study Group

Abstract: BACKGROUND AIMS Inflammatory bowel disease (IBD) scoring systems combine patient-reported data with physicians' observations to determine patient outcomes, but these systems are believed to have limitations. We used real-world data from a large IBD cohort in Switzerland to compare results between patients and healthcare professionals from scoring systems for Crohn's disease (CD) and ulcerative colitis (UC). METHODS We collected data from the Swiss IBD cohort, beginning in 2006, using 2453 reports for 1385 patients (52% female, 58% with CD). During office visits, physicians asked patients about signs and symptoms and recorded their answers (health care professional-reported outcomes). On a later date, patients received a questionnaire at home (independently of the medical visit), complete it, and sent it back to the data center. Patients also completed the short form 36 and IBD quality of life (QoL) questionnaires. We calculated Cohen's kappa ( ) statistics to assess the level of agreement in scores between patients and health care professionals (Δt between reports collected less than 2 months apart). We used Spearman correlation coefficients ( ) to compare general well-being (GWB) and QoL scores determined by patients vs health care professionals. Our primary aim was to investigate the overall and individual level of agreement on signs and symptoms reported by health care professionals vs patients. RESULTS The best level of agreement (although moderate) was observed for number of stools last week in patients with CD (= 0.47), and nocturnal diarrhea in patients with UC (= 0.52). Agreement was low on level of abdominal pain (= 0.31 for patients with CD and = 0.37 for patients with UC) and GWB (= 0.23 for patients with CD and = 0.26 for patients with UC). Patients reported less severe abdominal pain and worse GWB (CD) or better GWB (UC) than that determined by health care professionals. Patient self-rated GWB correlated with IBD quality of life (= 0.68 for patients with CD and = 0.70 for patients with UC) and SF-36 physical scores (= 0.55 for patients with CD and = 0.60 for patients with UC); there was no correlation between health care professional-rated GWB and QoL. CONCLUSIONS In a comparison of patient vs health care provider-reported outcomes in a Swiss IBD cohort, we found that health care professionals seem to misinterpret patients' complaints. Patients self-rated GWB correlated with QoL scores, indicating that reporting GWB in a single question is possible and relevant, but can vary based on how the data are collected.

DOI: https://doi.org/10.1016/j.cgh.2018.11.029

Pittet, Valérie E H; Maillard, Michel H; Simonson, Thomas; Fournier, Nicolas; Rogler, Gerhard; Michetti, Pierre; Swiss IBD Cohort Study Group (2019). Differences in Outcomes Reported by Patients With Inflammatory Bowel Diseases vs Their Health Care Professionals. Clinical Gastroenterology and Hepatology, 17(10):2050-2059.e1.
DOI: https://doi.org/10.1016/j.cgh.2018.11.029

Manuscript Number: CGH 18-01033

Title: Differences in Outcomes Reported by Patients With Inflammatory Bowel Diseases vs Their Health Care Professionals

Valérie Pittet[1], Michel H Maillard[2,4], Thomas Simonson[1], Nicolas Fournier[1], Gerhard Rogler[3], Pierre Michetti[2,4]

[1]Institute of Social & Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, Switzerland

[2]Division of Gastroenterology & Hepatology, Lausanne University Hospital, Lausanne, Switzerland

[3]Department of Gastroenterology & Hepatology, University Hospital Zurich, Zurich, Switzerland

[4]Crohn and Colitis Center, Gastroentérologie Beaulieu SA, Lausanne, Switzerland

[§]Corresponding author:

PD Valérie Pittet

Healthcare Evaluation Unit, Institute of Social & Preventive Medicine (IUMSP)

Route de la Corniche 10, CH-1010 Lausanne, Switzerland

Phone: +41 21 314 72 82, Email: Valerie.pittet@chuv.ch

SHORT TITLE: Reporting outcomes in real-life settings.

TOTAL WORD COUNT: 3842

CONTRIBUTORSHIP

Conception and design (VP, PM), statistical analysis or interpretation of data (VP, NF, TS), drafting of the manuscript (VP, TS), critical revision of the manuscript for significant

intellectual content (VP, MM, TS, NF, GR, PM), final approval of the version to be published (VP, MM, TS, NF, GR, PM), obtained funding (VP, GR).

COMPETING INTEREST

No author has any conflict of interest or financial ties relevant to the manuscript to disclose.

FUNDING

The study is supported by the Swiss National Science Foundation (SNSF) grants N°33CS30-148422 (Swiss IBD cohort study) to GR and 32473B-138498 (Appropriateness of care in IBD) to VP.

ABBREVIATIONS:

IBD: inflammatory bowel disease

CD: Crohn's disease

UC: ulcerative colitis

SIBDC: Swiss IBD cohort

HCP: healthcare professional

IBDQ: Inflammatory Bowel Disease Questionnaire

SF-36: Short Form (36) Health Survey Questionnaire

PCS: SF-36 Physical Component Summary score

MCS: SF-36 Mental Component Summary score

GWB: general well-being

PRO: patient-reported outcome

PROM: patient-reported outcome measure

CDAI: Crohn's Disease Activity Index

MTWAI: Modified Truelove and Witts activity index

QoL: quality of life

GI: gastrointestinal

EIM: extraintestinal manifestation

IQR: interquartile range

CI: confidence interval

RCT: randomized controlled trial

Abstract

Background & Aims: Inflammatory bowel disease (IBD) scoring systems combine patient-reported data with physicians' observations to determine patient outcomes, but these systems are believed to have limitations. We used real-world data from a large IBD cohort in Switzerland to compare results between patients and healthcare professionals' from scoring systems for Crohn's disease (CD) and ulcerative colitis (UC).

Methods: We collected data from the Swiss IBD cohort, beginning in 2006, using 2453 reports for 1385 patients (52% female, 58% with CD). During office visits, physicians asked patients about signs and symptoms and recorded their answers (health care professional-reported outcomes). On a later date, patients received a questionnaire at home (independently of the medical visit), complete it, and sent it back to the data center. Patients also completed the short form 36 and IBD quality of life (QoL) questionnaires. We calculated Cohen's kappa (κ) statistics to assess the level of agreement in scores between patients and health care professionals (Δt between reports collected less than 2 months apart). We used Spearman correlation coefficients (r) to compare general well-being (GWB) and QoL scores determined by patients vs health care professionals. Our primary aim was to investigate the overall and individual level of agreement on signs and symptoms reported by health care professionals vs patients.

Results: The best level of agreement (although moderate) was observed for number of stools last week in patients with CD (κ=0.47), and nocturnal diarrhea in patients with UC (κ=0.52). Agreement was low on level of abdominal pain (κ=0.31 for patients with CD and κ=0.37 for patients with UC) and GWB (κ=0.23 for patients with CD and κ=0.26 for patients with UC). Patients reported less severe abdominal pain and worse GWB (CD) or better GWB (UC) than that determined by health care professionals. Patient self-rated GWB correlated with IBD quality of life (r=0.68 for patients with CD and r=0.70 for patients with UC) and SF-36 physical scores (r=0.55 for patients with CD and r=0.60 for patients with UC); there was no correlation between health care professional-rated GWB and QoL.

Conclusion: In a comparison of patient vs health care provider-reported outcomes in a Swiss IBD cohort, we found that health care professionals seem to misinterpret patients' complaints. Patients self-rated GWB correlated with QoL scores, indicating that reporting GWB in a single question is possible and relevant, but can vary based on how the data are collected.

KEY WORDS: disease activity index; SF-36; PRO; management

Introduction

In inflammatory bowel diseases (IBD), treatment goals are based on clinical, endoscopic, biologic, and histologic endpoints[1, 2]. Signs and symptoms are currently assessed through disease activity scores, which consist of a mixture of patient-reported outcomes (PROs) and physician's observations, sometimes complemented by laboratory values or endoscopic evaluations. These scores have shown limitations in clinical trials[3], mainly because PROs were not constructed from patients' views about their symptoms and health-related quality of life, but from clinicians' interpretations of patients' views, thus introducing potential bias into PRO measures (PROMs). Indeed, PROMs are defined as measures that should ascertain patients' views about their life[4], to be reported directly by the patient without interpretation by a clinician or anyone else[5]. Pending the development of such ideal PROs, interim PROs have been proposed for use in clinical trials for Crohn's disease (CD)[6] and ulcerative colitis (UC)[7]. These PROs, however, were derived from retrospective data. They are thus not taken directly from patient-reported data, and their extraction has not been prospectively validated.

Tremendous work has been made during the past 5 years to redefine PROs,[8-12] both in general and more specifically to define a core set of PROs to be used in clinical trials[12-15]. PRO developments have also recently been reported in other chronic inflammatory or gastrointestinal (GI) diseases[16-18]. In IBD, the addition of PROs to clinical scores used by healthcare professionals (HCPs) can improve their ability to predict endoscopic remission[19]. Further, it is crucial for the promotion of patient empowerment programs to develop endpoints based on information reported by patients themselves and to use these endpoints in clinical trials[20, 21].

Defining symptom-based PROs is important because symptoms remain the central driver for patients in their decisions regarding care, potential treatment changes, diet adaptations, and healthcare help seeking. PROs will also be central for future self-management programs. Future management strategies favor tight control and monitoring of patients[22] to target preclinical signs

of relapse; the goal is to prevent disease progression and achieve better long-term outcomes. Thus, treatment and management decisions in IBD call for good translation of scores developed in trials into clinical practice, the goal being to use these scores to assess treatment effectiveness. In the first stage, the accuracy and practicability of assessing disease activity scores (including current PROs) should be determined through real-world data collected in routine settings by HCPs involved in regular follow-up visits. This would help to provide insight into the content and validity of the items generated from trials to be used in routine clinical practice[23].

Our primary aim was to investigate the overall and individual levels of agreement on the signs and symptoms of IBD reported by HCPs and patients. Our secondary aim was to study the feasibility of capturing patient-reported information on quality of life through a single simple question on general well-being used in activity indexes.

Methods

Patient population

The Swiss IBD cohort (SIBDC) is a national ongoing prospective cohort of incident and prevalent IBD cases that started in November 2006. Patients were recruited through their gastroenterologist. Recruitment was open to any gastroenterologist working in a tertiary center, regional hospital, or private practice. Patients' inclusion criteria were the following: being diagnosed for at least 2 months or having had at least 1 recurrence of symptoms, diagnosis confirmed by radiological or endoscopic and histological findings, living in Switzerland or having contracted Swiss healthcare insurance, and being treated for IBD in Switzerland.

Cohort data

We used data collected in the framework of the SIBDC to characterize patients and their disease. Clinical characteristics and treatments were collected during patients' yearly medical visits by gastroenterologists or trained study nurses (hereafter grouped as HCPs). Annual follow-up questionnaires were sent to patients to collect patient-reported data, including psychosomatic scores.

For descriptive purposes, the following were extracted from the SIBDC databases: gender, age, age at diagnosis, type of diagnosis (CD or UC), CD disease location (ileal, colonic, ileocolonic, upper GI only), UC extension (proctitis, left-sided colitis, pancolitis), history of extraintestinal manifestations (EIMs), CD history of perianal fistula, CD history of stenosis, and history of therapies. We also collected all items used to calculate disease activity scores, namely, the Crohn's Disease Activity Index (CDAI) for CD (number of soft stools last week, abdominal pain, general well-being [GWB], EIMs, antidiarrheal treatment, abdominal mass, hematocrit, body weight) and the Modified Truelove and Witts activity index (MTWAI) for UC (number of bowel movements per day, nocturnal diarrhea, bloody stools, fecal

incontinence, abdominal pain/cramping, GWB, antidiarrheal treatment, abdominal tenderness). Remission was defined as a CDAI score lower than or equal to 150 for CD, or an MTWAI score lower than or equal to 4 for UC. For the CDAI, patients recorded the number of stools retrospectively.

From the self-administered patient questionnaires, we extracted psychometric data on general and disease-related quality of life (QoL). We used the Short Form (36) Health Survey Questionnaire (SF-36), divided into 2 subscores, the Physical Component Summary score (PCS) and the Mental Component Summary score (MCS). The 32-question Inflammatory Bowel Diseases Questionnaire (IBDQ) was summed as a total score and divided into 4 subscores (bowel symptoms, systemic symptoms, emotional function, and social function). From 2012 to 2015, we also collected patient-reported information on signs, symptoms, and antidiarrheal treatment as follows: number of soft stools last week (number per day and total number), nocturnal diarrhea (yes, no), bloody stool frequency (continuous scale from 0% to 100%), fecal incontinence (yes, no), abdominal pain/cramps (continuous scale from 0 to 4, 0=never, 4=severe), GWB (continuous scale from 0 to 5, 0=very bad, 5=very well), antidiarrheal treatment (yes, no), fever in last 7 days (yes, no), surgery planned in the coming weeks (yes, no), EIMs. Patients were asked to complete all questions, regardless of their diagnosis. This decision was made from our previous observation that a non-negligible number of patients indicated another diagnosis (CD versus UC) than that established by their doctor. For this reason, we referred only to the diagnosis confirmed by the physician in the analyses.

The results presented hereafter were not stratified by disease activity (active versus remission). Stratification by disease activity was first performed, but did not show significant differences compared with non-stratified analysis. The number of cases for patients with active disease was 26 (4.8%) for CD and 31 (16.2%) for UC; thus, the number of cases was low, probably due to the difficulty in collecting information from the patient at the time of a flare in

a real-life situation. We found that kappa values were low regardless of whether patients had active or inactive disease and regardless of disease type, when the 2 reports occurred within a maximum of 1 month (κ=0.29 [remission] versus 0.31 [active disease] for CD, κ=0.15 [remission] versus 0.27 [active disease] for UC). From this observation, we decided to calculate kappas for the total number of observations, irrespective of (HCP-reported) disease activity. It is thus important to consider that the majority of patients were stable and in remission.

Statistical analysis

Descriptive analyses with numbers and percentages, or means and standard deviations, were used to characterize the study population.

We performed the analyses on the completed and comparable subset of items collected independently by HCPs and patients. As the patient self-reported questionnaire was completed at home, patients could freely decide on which date to complete it (i.e., immediately after receiving it or later). Therefore, we restricted the analyses to data collection made within a delta time (Δt) ranging from 0 to 2 months. Median time between the HCP report and the patient report was 26 days for CD (interquartile range [IQR] 11-41 days, range 0-59 days) and 26 days for UC (IQR 13-41 days, range 0-59 days). If a patient presented with several reports, all reports were included. HCP reports were spaced at least 1 year apart (HCP data collection is performed yearly). Agreement between patients and HCPs was assessed by using Cohen's kappa statistic. We stratified the analysis by diagnosis (CD and UC). Categorical variables were treated as ordinal, and a quadratic weight was applied to penalize more serious disagreements. For complex multi-item variables (EIMs), we computed a pooled kappa in accordance with the method of De Vries et al.[24]: we averaged the observed and expected probability of agreement over several subitems and computed a single pooled kappa by using these 2 quantities. A pooled kappa was also computed to summarize overall agreement over all examined variables. For

pooled kappa statistics, 95% confidence intervals (CIs) were obtained by using bootstrap resampling.

We calculated Spearman's rank correlation coefficients (rho) and 95% CIs to assess the direction and strength of the association between QoL scores and GWB. We used the whole patient population who had completed a self-reported questionnaire from 2012 to 2015 (N=2453) to calculate the correlation between patient-reported GWB and QoL scores, all information being collected in the same questionnaire. We restricted the Δt analyses to 0 to 2 months to calculate the correlation between HCP-reported GWB and QoL scores.

Statistics were produced by using STATA v.14.1 (STATA Corp., College Station, TX, USA) and R version 3.4.2 (2017-09-28).

Ethics approval

Ethics approval was obtained from the regional Swiss Ethics Committees in which cohort participants were enrolled (Commission d'éthique du Canton de Vaud/Protocol no. 33/06).

Results

Patient characteristics

At the time of data extraction, the SIBDC comprised 3556 patients. A response rate to the patient questionnaires over time of around 70%, as well as delta time restriction for comparison with HCP data, yielded a final sample size of 1385 patients for the present analysis, 800 with CD (58%) and 585 with UC (42%) (Table 1). About half were women (52.2%). Patients had a median age of 47 years and a median disease duration of 14 years. One third of patients with UC and two thirds of patients with CD received at least 1 anti-tumor necrosis factor agent during their disease course. A higher proportion of patients with CD had a history of EIM than did those with UC (63.1% versus 48.0%). Around two thirds of CD patients had ileal involvement, half experienced stenosis, and a quarter had a perianal fistula. One third of UC patients had extensive disease. Patients with CD had lower PCS/MCS subscores than did those with UC (48.6/46.2 versus 50.7/47.4, respectively). Overall, the largest deviation from the normal population was observed for mental scores (46.7 versus 50). IBDQ total scores and subscores were all lower for CD patients than for UC patients.

The following analyses were conducted on 2453 reports available from patients and/or HCPs. The number of reports that were evaluable per patient ranged from 1 to 4, with 60% of patients having 2 to 4 reports (Table 2). For 38% of those reports, the Δt between patients and HCP reports to the SIBDC was less than or equal to 2 months.

Agreement analyses

The following analyses were performed on the subgroup of patients and HCP reports made within Δt 0-2 months. This corresponded to 936 reports, 535 for CD patients and 401 for UC patients. In this analysis, 30% of all CD patients contributed with 2 to 3 observations; the proportion was similar among UC patients (Table 2).

Overall agreement

The overall (pooled) agreement between patients and HCP reports for all considered items was κ=0.35 (95% CI, 0.30–0.39) for CD and κ=0.31 (95% CI, 0.24–0.38) for UC (Figure 1).

Agreement on individual items

The best, although moderate, patient/HCP agreement was found for the number of soft stools reported in the last week (κ=0.47) in CD patients, and for nocturnal diarrhea (κ=0.52) and the presence of bloody stools (κ=0.41) in UC patients. The level of agreement was low and unrelated to type of diagnosis regarding abdominal pain (CD: κ=0.32, UC: κ=0.37) and GWB (CD: κ=0.23, UC: κ=0.26). CD patients and HCPs had low agreement on the use of antidiarrheal medications (κ=0.36), and UC patients had low agreement with HCPs on fecal incontinence (κ=0.19) and antidiarrheal treatment (κ=0.18). When exploring the direction of disagreement in CD, we found (Figure 2) that abdominal pain was often overestimated by clinicians compared with that reported by patients. Indeed, more than half of the patients for whom abdominal pain was reported as mild or moderate-to-severe by HCPs rated their pain as none. Figure 3 shows that abdominal pain was more frequently reported by UC patients as being absent or less severe than it was by HCPs. Of the patients for whom abdominal pain was reported as being mild or moderate-to-severe by HCPs, 50% to 60% rated it as none. GWB was frequently rated as being worse by CD patients than by HCPs. Indeed, when HCPs reported GWB as well, 32% of UC patients rated it as being below par and 21% as poor to terrible. In contrast, UC patients tended to rate GWB as being better than HCPs did. We indeed observed that, when HCPs reported GWB as being fair, 54% of patients rated it as excellent.

Correlation analyses

Correlation analyses with QoL scores were performed, first on the total number of reports available (N=2453) for patient-reported GWB, and then on the selected subgroup of patient/HCP reports made within Δt 0-2 months for HCP-reported GWB.

The mean (SD) value of patient-reported GWB was 3.8 (1.2) in CD patients and 3.9 (1.2) in UC patients (Table 2). HCP-reported GWB was rated as well/excellent in three quarters (76.9%) of CD patients and as well/excellent in one third (34.7%) of UC patients. Nearly half of UC patients were rated by HCPs as having poor/terrible GWB.

Spearman's rank correlation coefficients between GWB and QoL scores, Table 3, were all significant for the monotonic association, but the strength of the correlation varied considerably, depending on HCP and patient reports, type of disease, and QoL instrument. Overall, patient self-rated GWB was more strongly correlated with QoL scores than was HCP-reported GWB, and the correlation was better for UC patients than for CD patients. When reported by HCPs, GWB had little correlation (rho<0.3) with the MCS subscore (rho=-0.26) and the IBDQ emotional function subscore (rho=-0.28) in CD patients. Although still low (rho<0.5), the correlation was better for CD patients at Δt=1-2 months than at Δt=0-1 month. In UC patients, correlations between HCP-reported GWB and QoL scores were better when the time interval between both reports was the shortest. Patient self-rated GWB correlated moderately to highly with QoL subscores. The highest correlation was observed with the IBDQ total sum score (r=0.68 for CD, r=0.70 for UC) and the lowest with the MCS subscore (r=0.49 for CD, r=0.50 for UC). Overall, the correlation of patient self-rated GWB was better with IBDQ scores than with SF-36 scores.

Discussion

Herein we presented analyses performed on real-world data collected by HCPs and patients in the framework of the SIBDC. Using these data, we were able to assess the level of agreement between patients and HCPs on the signs and symptoms used for clinical activity score calculations. We found that the agreement between reports was in general low – high agreement corresponds to a kappa above 0.6 – especially for abdominal pain and GWB. We also assessed the feasibility of capturing information on patients' QoL by using a single question on GWB. We found that the correlation between answers to this single question and QoL scores was good when reported by patients but not when reported by HCPs. Patient self-rated GWB was highly correlated with the IBDQ total sum score, indicating that the answer to a single and simple question could indeed be used to reflect disease-related QoL, provided that the question is answered by patients in their own setting.

Two previous studies conducted by Bennebroek Evertsz et al. also made crude comparisons between HCP- and patient-reported items on disease activity items in CD[25] and UC[26]. For CD, they found lower agreement on abdominal pain (k=0.14) but higher on GWB (k=0.52) than we did in our study. In their study, CD patient-reported abdominal pain was never scored lower than that of HCPs, which was the opposite of our results. For UC, they found results similar to ours, i.e., the highest level of agreement for nocturnal diarrhea (k=0.67) and the presence of bloody stools (k=0.63), the lowest for fecal incontinence (k=0.26); the agreement for GWB (k=0.49) was moderate. A positive aspect of their studies was that both measures were taken within a short time frame, which may have favored slightly better agreement than occurred with the findings of our study. On the other hand, patient reports were collected in the doctor's setting, i.e., not independently in the patient's own environment. This may explain why, in the Bennebroek Evertsz et al. study, better agreement on GWB was not achieved, although the study conditions were good. Agreement is a measure of the difference

between HCP reports on patient signs or symptoms (i.e., made during a medical visit) and patient self-reports on the same parameters, but collected outside the medical practice (i.e., in the patient's living environment). Both reports were independent. When "hard quantitative" or "objective" measures are compared, we would expect the results to be the same because they are both reported by the patient, once in the HCP's environment and once in the patient's environment. In the case of subjective assessments, however, e.g., well-being, the double mechanism of reporting (by the patient) and interpretation of this reporting (by the physician/HCP) might lead to disagreement. Reasons for the disagreement might be linked to recall bias and socially desirable responding[27-29]. Response to reports can also be influenced by many factors, including stress and fear of medical visits among patients, busy teams and work overload among HCPs, or other societal and cultural factors. The question here was not to consider whose reporting was right or wrong, but to assess, on a similar question, potential sources of measurement errors that could be introduced by the context in which measures were taken, potentially leading to over- or underestimation of activity scores. This is especially important for items that have high weights in score calculations.

Recent studies have proposed the use of interim PROs in randomized controlled trials (RCTs)[6, 7] while awaiting the development of PROs by using standards provided by the US Food and Drug Administration. PRO-2 and 3, including GWB and abdominal pain for CD, were considered reasonable information reported by patients[13], hypotheses that are not confirmed by our data. However, these studies used patient information obtained during clinical trials, where the framework in which the data were collected differed from that of routine practice. Our data support the view that the use of PRO-2 or 3 is probably not transposable to real-life routine clinical practice or would require validation in this setting.

Abdominal pain is considered by patients to be an important item for assessment[13]. Our results suggest, however, that the setting for collection of this symptom modulates patients'

answers. Indeed, in both CD and UC, abdominal pain was rated lower when reported by patients than by HCPs. We hypothesize that this difference is related to an observer bias, since patients seem to systematically underrate abdominal pain when they are not in front of their physician. One reason might be a beneficial effect of the home environment on their feeling of pain compared with that at the doctor's practice. Alternatively, pain may be used to attract more of their doctor's attention to other symptoms. To improve the correlation of this measure, we could consider collecting additional information that may be associated with abdominal pain (i.e., disability, stress, sleep quality, fatigue)[30-32] because the question might be whether abdominal pain reflects a cause or a consequence of other symptoms.

We observed that patient self-rated GWB correlated much better with SF-36 and IBDQ scores than did HCP-reported GWB, with some limitations for the mental health component. This result suggests that it may be possible to assess disease-related QoL through a unique question on well-being. These findings further indicate that the context in which patients answer subjective questioning is crucial. GWB may be modified and less reliable when the patient is questioned by a doctor in a setting (time pressure, waiting room oversaturation) that can further alter the patient's response in unpredictable ways. We could have argued for a social desirability bias[33], but our results showed that the most favorable answer is not found in patients' home reports, but in the answers given in front of physicians, at least in CD. Our observation is, however, consistent with previous guidelines that reported a higher likelihood of face-to-face interviews to prompt positive reports about patient well-being than with self-administered questionnaires[33]. Among other factors that could be associated with GWB in patients is gender, which was shown to be associated with the perception of QoL[34], or potential variations in the clinical interpretation of subjective measures of disease activity.

One way to move highly subjective PROs outside the clinical setting would be to use remote or mobile tools. Several attempts have already been made in this direction, allowing

shared data collection[35, 36]. Telemedicine systems that facilitate the sharing of information and feedback have proven to be useful in disease activity monitoring[23, 37]. Therefore, further attempts in this direction should be tested. For example, GWB could be completed by the patient at home on request of the treating physician, e.g., 24 hours before the medical appointment, so that it could be used during the consultation.

The strength of our study was the ability to use real-life data collected through an ongoing cohort study. This allowed us to investigate how data, which should be collected in the same way as in clinical trials for decision making, could be collected in the reality of current practice. The other strength of our study was the unique collection of patient-reported data that was done in parallel with the collection of clinical data.

Our study has some limitations. In performing agreement analyses, the ideal interval between clinic follow-up visit and questionnaire participation should be as short as possible. In our study, this limitation was compensated for by the opportunity to have independent measures made by the patient and the HCP. This allowed us to investigate differences according to the context in which data collection was made. Moreover, the large majority of patients in our study were in remission; therefore, we can assume a low likelihood of major changes in the patients' symptoms within a 2-month time interval. Nevertheless, a previous study showed that patients may experience more flares than those discussed with the physician[38]. Therefore, we could not ensure that all patients were in the disease activity stage, as reflected by the sum score of the HCP-reported activity index.

To bridge the gap between clinical research and practice, we need simple PROs that can be used in a standardized way both in RCTs and in the real-world clinical context. Our results show that it is possible to capture information reflecting patients' signs, even on the basis of single questions. This condition is essential for routine clinical practice. However, the differences we observed require further research in 2 directions: first, in PRO research, which

is already the scope of ongoing studies, and second, in assessing the feasibility and practicability of methods for collecting such data, particularly with respect to subjective measures in which the context of data collection, a proxy of known and unknown confounders, appears to have an impact.

REFERENCES

1. D'Haens G, Sandborn WJ, Feagan BG, et al. A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis. Gastroenterology 2007;132:763-86.
2. Sandborn WJ, Feagan BG, Hanauer SB, et al. A review of activity indices and efficacy endpoints for clinical trials of medical therapy in adults with Crohn's disease. Gastroenterology 2002;122:512-30.
3. Levesque BG, Sandborn WJ, Ruel J, et al. Converging goals of treatment of inflammatory bowel disease from clinical trials and practice. Gastroenterology 2015;148:37-51 e1.
4. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346:f167.
5. Weldring T, Smith SM. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). Health Serv Insights 2013;6:61-8.
6. Khanna R, Zou G, D'Haens G, et al. A retrospective analysis: the development of patient reported outcome measures for the assessment of Crohn's disease activity. Aliment Pharmacol Ther 2015;41:77-86.
7. Jairath V, Khanna R, Zou GY, et al. Development of interim patient-reported outcome measures for the assessment of ulcerative colitis disease activity in clinical trials. Aliment Pharmacol Ther 2015;42:1200-10.
8. Achleitner U, Coenen M, Colombel JF, et al. Identification of areas of functioning and disability addressed in inflammatory bowel disease-specific patient reported outcome measures. J Crohns Colitis 2012;6:507-17.
9. Williet N, Sarter H, Gower-Rousseau C, et al. Patient-reported Outcomes in a French Nationwide Survey of Inflammatory Bowel Disease Patients. J Crohns Colitis 2017;11:165-174.
10. IsHak WW, Pan D, Steiner AJ, et al. Patient-Reported Outcomes of Quality of Life, Functioning, and GI/Psychiatric Symptom Severity in Patients with Inflammatory Bowel Disease (IBD). Inflamm Bowel Dis 2017;23:798-803.
11. Allen PB, Kamm MA, Peyrin-Biroulet L, et al. Development and validation of a patient-reported disability measurement tool for patients with inflammatory bowel disease. Aliment Pharmacol Ther 2013;37:438-44.
12. Williet N, Sandborn WJ, Peyrin-Biroulet L. Patient-reported outcomes as primary end points in clinical trials of inflammatory bowel disease. Clin Gastroenterol Hepatol 2014;12:1246-56 e6.
13. Bojic D, Bodger K, Travis S. Patient Reported Outcome Measures (PROMs) in Inflammatory Bowel Disease: New Data. J Crohns Colitis 2017;11:S576-S585.
14. Kim AH, Roberts C, Feagan BG, et al. Developing a Standard Set of Patient-Centred Outcomes for Inflammatory Bowel Disease-an International, Cross-disciplinary Consensus. J Crohns Colitis 2018;12:408-418.
15. Ma C, Panaccione R, Fedorak RN, et al. Heterogeneity in Definitions of Endpoints for Clinical Trials of Ulcerative Colitis: A Systematic Review for Development of a Core Outcome Set. Clin Gastroenterol Hepatol 2017.
16. Janta I, Naredo E, Martinez-Estupinan L, et al. Patient self-assessment and physician's assessment of rheumatoid arthritis activity: which is more realistic in remission status? A comparison with ultrasonography. Rheumatology (Oxford) 2013;52:2243-50.
17. Spiegel BM, Hays RD, Bolus R, et al. Development of the NIH Patient-Reported Outcomes Measurement Information System (PROMIS) gastrointestinal symptom scales. Am J Gastroenterol 2014;109:1804-14.

18. Schoepfer AM, Straumann A, Panczak R, et al. Development and validation of a symptom-based activity index for adults with eosinophilic esophagitis. Gastroenterology 2014;147:1255-66 e21.
19. Zittan E, Kabakchiev B, Kelly OB, et al. Development of the Harvey-Bradshaw Index-pro (HBI-PRO) Score to Assess Endoscopic Disease Activity in Crohn's Disease. J Crohns Colitis 2017;11:543-548.
20. Anderson R.M. FMM. Using the empowerment approach to help patients change behavior. American Diabetes Association. In: Anderson B.J. RRR, Editors, ed. Practical Psychology for Diabetes Clinicians, 2nd Edition: American Diabetes Association, 2002.
21. US FOOD AND DRUG ADMINISTRATION. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Federal Register 2009;74:65132-133.
22. Colombel JF, Narula N, Peyrin-Biroulet L. Management Strategies to Improve Outcomes of Patients With Inflammatory Bowel Diseases. Gastroenterology 2017;152:351-361 e5.
23. de Jong MJ, Huibregtse R, Masclee AAM, et al. Patient-Reported Outcome Measures for Use in Clinical Trials and Clinical Practice in Inflammatory Bowel Diseases: A Systematic Review. Clin Gastroenterol Hepatol 2017.
24. De Vries H, Elliott MN, Kanouse DE, et al. Using pooled kappa to summarize interrater agreement across many items. Field Methods 2008;20:272-282.
25. Bennebroek Evertsz F, Hoeks CC, Nieuwkerk PT, et al. Development of the patient Harvey Bradshaw index and a comparison with a clinician-based Harvey Bradshaw index assessment of Crohn's disease activity. J Clin Gastroenterol 2013;47:850-6.
26. Bennebroek Evertsz F, Nieuwkerk PT, Stokkers PC, et al. The patient simple clinical colitis activity index (P-SCCAI) can detect ulcerative colitis (UC) disease activity in remission: a comparison of the P-SCCAI with clinician-based SCCAI and biological markers. J Crohns Colitis 2013;7:890-900.
27. Alexander GC, Casalino LP, Meltzer DO. Patient-physician communication about out-of-pocket costs. JAMA 2003;290:953-8.
28. Matthews BA, Baker F, Spillers RL. How true is true? Assessing socially desirable response bias. Quality & Quantity 2003;37:327-335.
29. Moshagen M, Musch J, Ostapczuk M, et al. Reducing socially desirable responses in epidemiologic surveys: an extension of the randomized-response technique. Epidemiology 2010;21:379-82.
30. Jelsness-Jorgensen LP, Frigstad SO, Moum B, et al. Pain may be an important factor to consider in inflammatory bowel disease patients troubled by fatigue. United European Gastroenterol J 2017;5:687-693.
31. Norton C, Czuber-Dochan W, Artom M, et al. Systematic review: interventions for abdominal pain management in inflammatory bowel disease. Aliment Pharmacol Ther 2017;46:115-125.
32. Sweeney L, Moss-Morris R, Czuber-Dochan W, et al. Systematic review: psychosocial factors associated with pain in inflammatory bowel disease. Aliment Pharmacol Ther 2018;47:715-729.
33. Guidelines on Measuring Subjective Well-being. In: OECD, ed, 2013.
34. Tribbick D, Salzberg M, Connell W, et al. Differences Across Illness Perceptions in Inflammatory Bowel Disease and Their Relationships to Psychological Distress and Quality of Life. Gastroenterol Nurs 2017;40:291-299.

35. Marin-Jimenez I, Nos P, Domenech E, et al. Diagnostic Performance of the Simple Clinical Colitis Activity Index Self-Administered Online at Home by Patients With Ulcerative Colitis: CRONICA-UC Study. Am J Gastroenterol 2016;111:261-8.
36. Van Deen WK, van der Meulen-de Jong AE, Parekh NK, et al. Development and Validation of an Inflammatory Bowel Diseases Monitoring Index for Use With Mobile Health Technologies. Clin Gastroenterol Hepatol 2016;14:1742-1750 e7.
37. de Jong MJ, van der Meulen-de Jong AE, Romberg-Camps MJ, et al. Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. Lancet 2017;390:959-968.
38. Rubin DT, Siegel CA, Kane SV, et al. Impact of ulcerative colitis from patients' and physicians' perspectives: Results from the UC: NORMAL survey. Inflamm Bowel Dis 2009;15:581-8.

Figure 1: Cohen's Kappa (κ), 95% confidence intervals, for patient/HCP agreement on disease activity score items. Gray colors indicate ranges for κ interpretation (0-0.2: very low agreement, 0.2-0.4: low, 0.4-0.6: moderate, 0.6-0.8: high, 0.8-1: perfect).

Figure 2: Differences between CD patient/HCP ratings on abdominal pain and general well-being.

Figure 3: Differences between UC patient/HCP ratings on abdominal pain and general well-being.

TABLES

Table 1: Patients' clinical and psychometric characteristics

	CD	UC	Total
Number of patients	800 (57.8)	585 (42.2)	1385
Gender			
Male	367 (45.9)	295 (50.4)	662 (47.8)
Female	433 (54.1)	290 (49.6)	723 (52.2)
Age at diagnosis (median, range)	26, 3 – 78	31, 3 – 79	28, 3 – 79
Age (median, range)	47, 18 – 94	48, 19 – 89	47, 18 – 94
Years of disease duration (median, range)	15, 1 – 57	13, 1 - 54	14, 1 - 57
Therapy history			
5-ASA	492 (61.5)	561 (95.9)	1053 (76.0)
Steroids	701 (87.6)	494 (84.4)	1195 (86.3)
Immunomodulators	674 (84.3)	394 (67.4)	1068 (77.1)
Anti-TNF	498 (62.3)	211 (36.1)	709 (51.2)
History of EIM	505 (63.1)	281 (48.0)	786 (56.8)
History of Perianal Fistula	214 (26.7)		
History of Stenosis	381 (47.6)		
CD disease location			
L1 (ileal)	247 (30.9)		
L2 (colonic)	274 (24.3)		
L3 (ileocolonic)	243 (30.4)		
L4 (Upper GI only)	22 (2.8)		
Unknown/unclear	14 (1.8)		
UC disease extension			
Pancolitis		204 (34.9)	
Left-sided colitis		250 (42.7)	
Proctitis		123 (21.0)	
Unknown/Unclear		8 (1.4)	
SF36 (mean, SD)			
Physical Component Summary (PCS)	48.6 (9.4)	50.7 (8.4)	49.5 (9.1)
Mental Component Summary (MCS)	46.2 (11.3)	47.4 (9.9)	46.7 (10.7)
IBDQ (mean, SD)			
Bowel subscore	57.3 (9.6)	59.0 (9.8)	58.0 (9.8)
Systemic subscore	25.2 (6.1)	26.0 (5.9)	25.5 (6.0)
Emotional subscore	66.2 (12.7)	67.7 (11.7)	66.8 (12.3)
Social subscore	30.7 (5.7)	31.7 (5.4)	31.1 (5.6)
Total	179.5 (30.5)	184.3 (29.3)	181.5 (30.1)

Table 2: Description of the number of reports available from 2012 to 2015, the Δt between patient/HCP reports, and patient/HCP overall rating scores on GWB.

	CD	UC	Total
Number of Patient-reports available per patient			
1	322 (40.3)	254 (43.4)	576 (41.6)
2	321 (40.1)	237 (40.5)	558 (40.3)
3-4	157 (19.6)	94 (16.1)	251 (18.2)
Δt between HCP/Patient reports			
0-1 month	318 (22.1)	224 (22.1)	542 (22.1)
1-2 months	217 (15.1)	177 (17.5)	394 (16.1)
Total reports (patients)with Δt 0-2 months	*535 (398)*	*401 (310)*	*936 (708)*
2-4 months	526 (36.6)	355 (35.1)	881 (35.9)
4-6 months	378 (26.3)	258 (25.4)	636 (25.9)
Total	*1439 (100.0)*	*1014 (100.0)*	*2453 (100.0)*
Patient self-rated single question on GWB (mean, SD)	3.8 (1.2)	3.9 (1.2)	3.8 (1.2)
HCP-reported GWB for Δt <= 2months			
Well / Excellent	392 (76.9)	132 (34.7)	524 (58.8)
Below par / Fair	87 (17.1)	72 (18.9)	159 (17.9)
Poor-terrible / Poor	31 (6.1)	177 (46.5)	208 (23.3

Table 3: Spearman's rank correlation coefficient rho (95% confidence intervals) between ratings on GWB and QoL scores.

	GWB rated by HCP			GWB rated by P
	0-1 months	**1-2 months**	**0-2 months**	
CROHN'S DISEASE				
SF-36 PCS	0.30 (0.18 to 0.41)	0.47 (0.36 to 0.59)	0.38 (0.30 to 0.45)	0.55 (0.51 to 0.59)
SF-36 MCS	0.23 (0.11 to 0.36)	0.27 (0.15 to 0.40)	0.26 (0.17 to 0.35)	0.49 (0.45 to 0.54)
IBDQ Bowel	0.30 (0.20 to 0.40)	0.42 (0.31 to 0.54)	0.35 (0.28 to 0.43)	0.59 (0.55 to 0.63)
IBDQ Systemic	0.28 (0.17 to 0.40)	0.42 (0.30 to 0.55)	0.35 (0.28 to 0.43)	0.65 (0.61 to 0.68)
IBDQ Emotional	0.23 (0.12 to 0.34)	0.34 (0.22 to 0.47)	0.28 (0.20 to 0.36)	0.60 (0.56 to 0.64)
IBDQ Social	0.23 (0.11 to 0.34)	0.41 (0.28 to 0.53)	0.31 (0.22 to 0.39)	0.58 (0.54 to 0.62)
IBDQ Total	0.29 (0.19 to 0.40)	0.43 (0.31 to 0.55)	0.35 (0.28 to 0.43)	0.68 (0.64 to 0.71)
ULCERATIVE COLITIS				
SF-36 PCS	0.32 (0.19 to 0.45)	0.30 (0.16 to 0.44)	0.31 (0.21 to 0.41)	0.60 (0.55 to 0.65)
SF-36 MCS	0.34 (0.20 to 0.47)	0.18 (0.03 to 0.33)	0.22 (0.17 to 0.37)	0.50 (0.45 to 0.56)
IBDQ Bowel	0.31 (0.18 to 0.45)	0.30 (0.14 to 0.46)	0.31 (0.21 to 0.40)	0.64 (0.59 to 0.68)
IBDQ Systemic	0.41 (0.29 to 0.53)	0.31 (0.15 to 0.46)	0.36 (0.27 to 0.46)	0.62 (0.58 to 0.67)
IBDQ Emotional	0.41 (0.30 to 0.53)	0.30 (0.14 to 0.48)	0.36 (0.27 to 0.45)	0.61 (0.56 to 0.65)
IBDQ Social	0.34 (0.22 to 0.47)	0.34 (0.21 to 0.52)	0.35 (0.25 to 0.45)	0.60 (0.55 to 0.65)
IBDQ Total	0.43 (0.31 to 0.54)	0.34 (0.19 to 0.34)	0.39 (0.30 to 0.48)	0.70 (0.66 to 0.74)

Interpretation of the rho value: $|rho|<0.3$ indicate little to no correlation, $0.5<|rho|<0.3$ low correlation, $0.7<|rho|<0.5$ moderate correlation, $|rho|>0.7$ high correlation.